Машинный перевод без параллельного текста

Мазуров Михаил

October 2018

1. Abstract

Рассматривается задача машинного перевода с одного языка на другой. Подготовка корпуса параллельных предложений является ресурсоемкой задачей. Предлагается метод построения нейросетевой модели, которая сможет переводить фразы из одного языка в другой и без использования корпуса параллельных предложений. Метод основан на модели Seq2Seq. Предлагается отображать предложения из двух языков в общее векторное пространство. Вычислительный эксперимент проводится на паре языков "русский-украинский".

2. Введение

Статья посвящена проблеме машинного перевода без параллельных пар предложений, данная задача возникает при построении систем перевода для редких пар языков и имеет ряд решений, основанных на нейросетевом машинном переводе. Основной идеей подхода является оптимизация Seq2Seq модели при сопоставлении векторных пространств слов на паре языков "русский-украинский".

Для оптимизации системы машинного перевода использованы рекурентные нейронные сети для кодировки и декодировки с векторным пространством в нужный нам язык, данный метод позволяет итеративно улучшать модель перевода, использую в качестве вспомогательного переводчика модель, полученную на предыдущих эпохах оптимизации. Этот подход протестирован на паре языков "английский - французски", показав хорошие результаты на уровне 27 BLEU (Bahdanau, Cho, & Bengio, 2014)

Также используются attention-механизмы, которые показывают свою состоятельность, давая улучшение в качесте машинного перевода (Luong, Pham, & Manning, 2015)

Для предотвращения с переобучением кодировщиков используется механизм зашумления предложений, позволяющий восстанавливать исходное предложение из зашумленного представления. Предлагается использовать автокодировщик, который будет учиться убирать шум в предложениях без перевода на другой язык.(Kim & Ney, 2018) Вычислительный эксперимент

ставится на паре “русский-украинский”. В качестве метрики качества перевода выступает метрика BLEU.

3. Постановка задачи

Мы работаем с двумя выборками непараллельных предложений на разных языках: $\mathfrak{D}^{\text{src}} = [\mathbf{s}_1^{\text{src}}, \dots, \mathbf{s}_{m_{\text{src}}}^{\text{src}}]$. и $\mathfrak{D}^{\text{tgt}} = [\mathbf{s}_1^{\text{tgt}}, \dots, \mathbf{s}_{m_{\text{tgt}}}^{\text{tgt}}]$.. Данные выборки будем использовать для тренировки модели машинного перевода. Для проверки алгоритма будем использовать валидационную выборку параллельных предложений $\mathfrak{D}^{valid} = \{(\mathbf{s}_1^{\text{src}}, \mathbf{s}_1^{\text{tgt}}), \dots, (\mathbf{s}_{m_{\text{valid}}}, \mathbf{s}_{\text{valid}})\}$.. Модель состоит из кодировщика $\mathbf{f}$ и декодировщика $\mathbf{g}$ для каждого языка, которые отображают предложения в общее векторное пространство и обратно.
Итоговая задача - уменьшить ошибку на валидационной выборке по трем видам оценки: по парам предложений, пословно внутри каждого предложения и ошибку предсказания. Также для противодействия переобучения введем функцию зашумления предложений - σ. Вводим три функционала ошибки, которые хотим минимизировать.

- Функционал ошибки по парам предложений

$$L_{\text{AE}} = ||\mathbf{g}(\mathbf{f}(\sigma(x))) - x||$$

- Функционал ошибки по словам внутри каждого предложения

$$L_{\text{TR}} = ||\mathbf{g}(\mathbf{f}(\mathbf{g}^{-1}(\mathbf{f}(x))) - x||$$

- Если есть модель , различающая скрытые представления векторов предложений из двух языков

$$L_{ADV} = \log \mathbb{P}(\mathbf{f}_{\text{lang}} = 1|\mathbf{g}(x)) + \log \mathbb{P}(\mathbf{f}_{\text{lang}} = 2|\mathbf{g}(y))$$

- Тогда итоговая функция ошибки будет принимать такой вид:

$$L = a * L_{AE} + b * L_{TR} + c * L_{ADV}$$

4. Базовый алгоритм

4.1. Получение слабого перевода

Сгенерирован словарь пар слов на основе смежных слов в двух парах словарей "русско-английский"и "англо-украинский"из (Conneau, Lample, Ranzato, Denoyer, & Jégou, 2017). Данные словари можно считать реальными выборками. Далее строится алгоритм, который делит предложения для перевода на слова и пословно ищет значения в сгенерированном словаре, если значения не находятся, то просто возвращается оригинальное слово.

Оценка работы алгоритма проводится с помощью BLEU-метрики. Результаты работы базового алгоритма на реальной выборке из субтитров к одному фильму на паре языков "русский-украинский"оценен:

$$BLEU = 10.86, 27.2/12.9/7.8/5.1(BP = 1.000, ratio = 1.010, hyp\_len = 3308640, ref\_len = 3275742)$$

4.2. Альтернативный метод

Берутся два словаря из (Conneau et al., 2017), соотносящие каждому слову некоторое векторное представление, причем векторное пространство общее и синтетически размечено так, что для слов, являющимися реальными переводами друг друга в разных языках, векторное представление приблизительно одинаковое. Мы строим алгоритм, переводящий слово в векторное пространство и в нем ищем наиболее близкие векторные представления из другого языка. Расстояние оценивается по L2 норме.

Список литературы

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Kim, Y., & Ney, J. G. H. (2018). Improving unsupervised word-by-word translation with language model and denoising autoencoder.

Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.